U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Data Warehouses

Data Warehouse for National Exams

Master in Data Science and Engineering

2º Semester - 2024/2025

João Soares – up202105364

João Viterbo Vieira – up202107689

Manuel Maria – up202108874

Professor Doutor: Gabriel David

Index

Index 2

1. Introduction 4

1.1 Subject 4

1.2 Goals 5

2. Planning 6

2.1 Dimensional Bus Matrix 6

2.1.1 Four-step Method 6

2.1.2 Dimensions 7

2.1.3 Facts 7

2.2 Facts Dictionary 8

2.2.1 Grades 8

2.2.2 Analytics_Grades 9

2.2.3 Analytics_Grades_School 10

2.2.4 Analytics_Grades_District 10

2.2.5 Analytics_Grades_Demographic 11

2.2.6 Analytics_Grades_Course 12

2.3 Dimensions Dictionary 12

2.3.1 Exam Dimension 12

2.3.2 Year Dimension 13

2.3.3 Phase Dimension 13

2.3.4 Course Dimension 13

2.3.5 School Dimension 14

2.3.6 Type Dimension 14

2.3.7 Subtype Dimension 14

2.3.8 District Dimension 15

2.3.9 Students Demographic Dimension 15

3. Dimensional Data Model 15

3.1 Overall Architecture 16

3.2 Star Schema Design 16

3.3 Fact Tables and Measures 16

3.4 Dimension Design Considerations 17

4. ETL and Data Sources 17

4.1 Extraction 18

4.2 Transformation 19

4.3 Loading 20

5. Query and Data Analysis 21

5.1 Query Performance Comparison 21

5.1.1 Average Exam Grade by Year and Subject..21
5.1.2 Exam Performance by School Type (Public vs. Private)....................................22
5.1.3 Regional Performance Trends Over Time... 22
5.1.4 Compare Average Grades by District.. 23
5.1.5 Compare School Performance.. 23
5.2 Results and Analysis...24
6. Visualizations.. 24
6.1 Dashboard Architecture..25
6.2 General Dashboard Analysis..25
6.3 Schools Dashboard Analysis..26
6.4 Districts Dashboard Analysis..26
6.5 Benefits of Dimensional Model for Visualization and Data Analysis.......................... 27
7. Conclusions..28
7.1 Pros and Cons of Data Warehouses vs Operational Databases...................................28
7.1.1 Advantages of Data Warehouses...28
7.1.2 Limitations of Data Warehouses..29
7.1.3 When Operational Databases Excel...29
7.2 Future Developments...29
7.2.1 Cloud-Native Data Warehousing... 30
7.2.2 Data Lakehouse Architecture... 30
7.2.3 AI/ML Integration.. 30
7.2.4 Data Mesh Architecture..30
8. References...31
9. Appendix..31

1. Introduction

In the digital age, educational institutions generate vast amounts of data that, when properly organized and analyzed, can provide valuable insights into student performance, curriculum effectiveness, and educational trends. However, operational databases, while efficient for day-to-day transactions, often lack the structure necessary for complex analytical queries and comprehensive reporting. This project seeks to overcome this limitation by designing and implementing a data warehouse for Portuguese national examination results from 2013 and 2014. Its flexible structure is built to seamlessly incorporate future data, enabling the development of a valuable longitudinal database.

The data warehouse approach offers significant advantages over traditional operational databases, including optimized query performance, historical data preservation, and enhanced analytical capabilities. By transforming our operational model into a dimensional model with a star schema design, we can facilitate multidimensional analysis of examination data, enabling education stakeholders to make more informed decisions.

1.1 Subject

This project focuses on the transformation and analysis of Portuguese national examination data from 2013 and 2014. The operational database contains approximately 1,000,000 records of examination results across various subjects, schools, and regions. The current operational model (Figure 1 in Appendix) consists of interconnected tables such as *tblHomologa_2014*, *tblEscolas*, *tblExames*, *tblCursos*, and several reference tables that store information about districts, municipalities, and school types.

These national exams are conducted across multiple districts, schools, courses, and academic years, offering a valuable opportunity for multidimensional analysis. To enhance accessibility and facilitate data-driven insights, our project transforms the complex relational structure of the operational database into a more analysis-friendly star schema. The core of this transformation is the creation of fact tables that store key performance metrics such as average grade, minimum and maximum scores and median results. The model connects these metrics to dimension tables including *dim_district*, *dim_school*, *dim_exam*, *dim_year*, *dim_course*, and *dim_studentDemographi*c, allowing for sophisticated analyses such as regional performance comparisons, year-over-year trends, course difficulty evaluation, and demographic correlations.

Although our current implementation is based on 2013 and 2014 examination data, our dimensional model design significantly reduces the time required to retrieve and analyze data. Additionally, the data warehouse is structured to incorporate future examination records, transforming it into a longitudinal database that will become increasingly valuable for historical and predictive educational analysis.

1.2 Goals

The primary objective of our project is to design and implement a data warehouse that facilitates efficient querying and analysis of the national examination data. Specifically, this project seeks to:

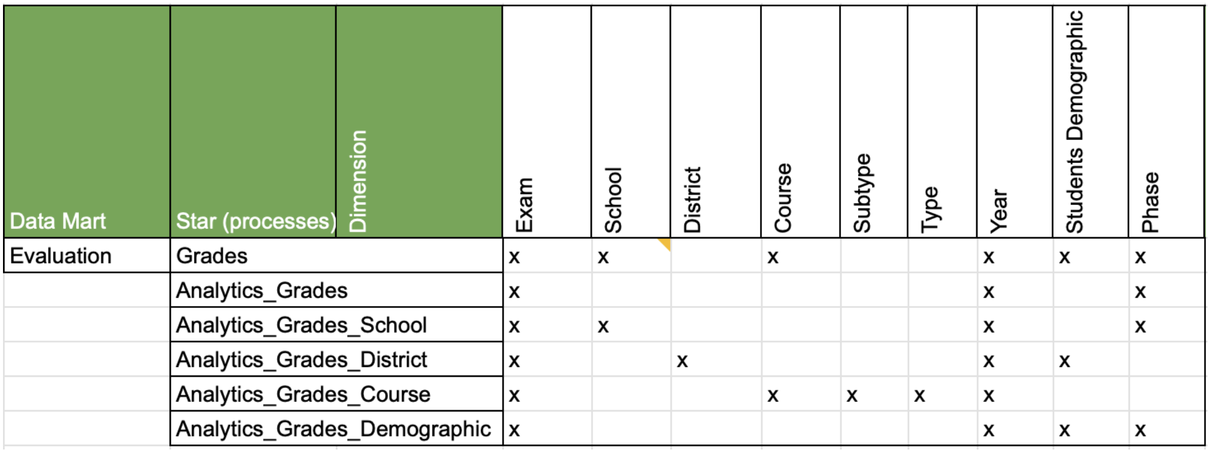
1. Transform our existing operational database into a dimensional model using a star schema design that complies with the assignment requirements
2. Develop a comprehensive ETL (Extract, Transform, Load) process to migrate data from the operational database to the data warehouse, ensuring data quality and consistency.
3. Create meaningful analytical queries that provide insights into examination performance across different dimensions such as time, subject, region, and school, highlighting differences in query performance between models.
4. Develop visualizations that illustrate key trends and patterns in the examination data, such as regional performance disparities, course difficulty analysis, and demographic factors.
5. Critically evaluate the advantages and shortcomings of the dimensional model compared to the relational model, focusing on query performance and data accessibility.

By achieving these goals, we aim to demonstrate the advantages of data warehousing in enhancing the efficiency and performance of educational data analysis. Through this project, we seek to highlight how a dimensional model improves query speed, data accessibility, and analytical capabilities compared to a traditional relational model.

2. Planning

2.1 Dimensional Bus Matrix

The bus matrix presented in the image above is a crucial planning tool for our data warehouse design, illustrating the relationship between our business processes, represented as fact tables, and the various dimensions through which we want to analyze the examination data. This matrix helps guide the dimensional model design and ensures consistency across the data warehouse.



Data Mart	Star (processes)	Dimension	Exam	School	District	Course	Subtype	Type	Year	Students Demographic	Phase
Evaluation	Grades		x	x		x			x	x	x
	Analytics_Grades		x						x		x
	Analytics_Grades_School		x	x					x		x
	Analytics_Grades_District		x		x				x	x	
	Analytics_Grades_Course		x			x	x	x	x		
	Analytics_Grades_Demographic		x						x	x	x

2.1.1 Four-step Method

In designing the fact tables and dimensions, we used the **four-step method**, which guides the development of an effective dimensional model. This method ensures that our data warehouse will be optimized for querying, performance, and scalability. The four steps are:

1. **The Data Mart:** We started by focusing on a single data mart for evaluation data, which simplifies our initial design and allows for a more targeted and efficient warehouse structure.
2. **Fact Table Granularity:** We carefully determined the granularity of our fact tables, ensuring that the data captured is both detailed enough to support in-depth analysis and flexible enough for future requirements.
3. **The Dimensions:** The dimensions were chosen based on their relevance to the data mart's business processes. They provide different perspectives from which we can analyze the examination data.

4. **The Facts:** The fact tables themselves were designed to capture the necessary measures at each level of granularity, ensuring that the data is appropriately aggregated and can be compared across different dimensions.

2.1.2 Dimensions

The dimensional model maps these business processes to nine key dimensions, which provide the various perspectives necessary for analyzing the examination data:

- ***Exam:*** Information about examination subjects
- ***Year:*** Temporal dimension for academic years
- ***Phase:*** Examination phases
- ***School:*** School information with geographic hierarchy
- ***Course:*** Course information with curriculum hierarchy
- ***StudentDemographic:*** Student demographic information
- ***Type:*** Course type classification
- ***SubType:*** Course subtype classification
- ***District:*** Regional classification by District

These dimensions enable comprehensive analysis, allowing the data to be explored across different contexts, such as by school, region, or student.

2.1.3 Facts

Our data warehouse consists of a **single data mart** focused on *Evaluation,* which contains six star schema processes. Each fact table represents a specific business process and its associated measures, designed to support different types of analysis:

- ***Grades:*** This is our primary fact table, containing individual student examination results. It connects with seven dimensions, which are Exam, School, Course, Year, Students Demographic, and Phase.
- ***Analytics_Grades:*** An aggregated fact table that focuses purely on exam performance across years and phases, allowing for time-based analysis of overall exam performance without geographic or institutional context.
- ***Analytics_Grades_School:*** This aggregated fact table provides school-level exam performance, connecting to Exam, School, Year, and Phase dimensions. This enables comparison of performance metrics across different schools over time.

- ***Analytics_Grades_District:*** Focused on district-level geographic analysis, this fact table connects Exam, District, Year, and Students Demographic dimensions, allowing for regional performance comparisons and demographic analysis within regions.
- ***Analytics_Grades_Course:*** This fact table enables curriculum-based analysis, connecting to Exam, Course, Subtype, Type, and Year dimensions. It facilitates performance comparison across different course types.
- ***Analytics_Grades_Demographic:*** This fact table focuses solely on demographic analysis, connecting Exam, Year, Students Demographic, and Phase dimensions to reveal how different demographic groups perform across exams.

These fact tables support multi-dimensional analysis by capturing essential data at different granularities, enabling insightful examination performance analysis from various perspectives.

Finally, by using conformed dimensions, which ensure consistency across all fact tables, and normalized facts, which help maintain data integrity and reduce redundancy, we ensure that our data warehouse is both accurate and reliable for cross-process analysis.

2.2 Facts Dictionary

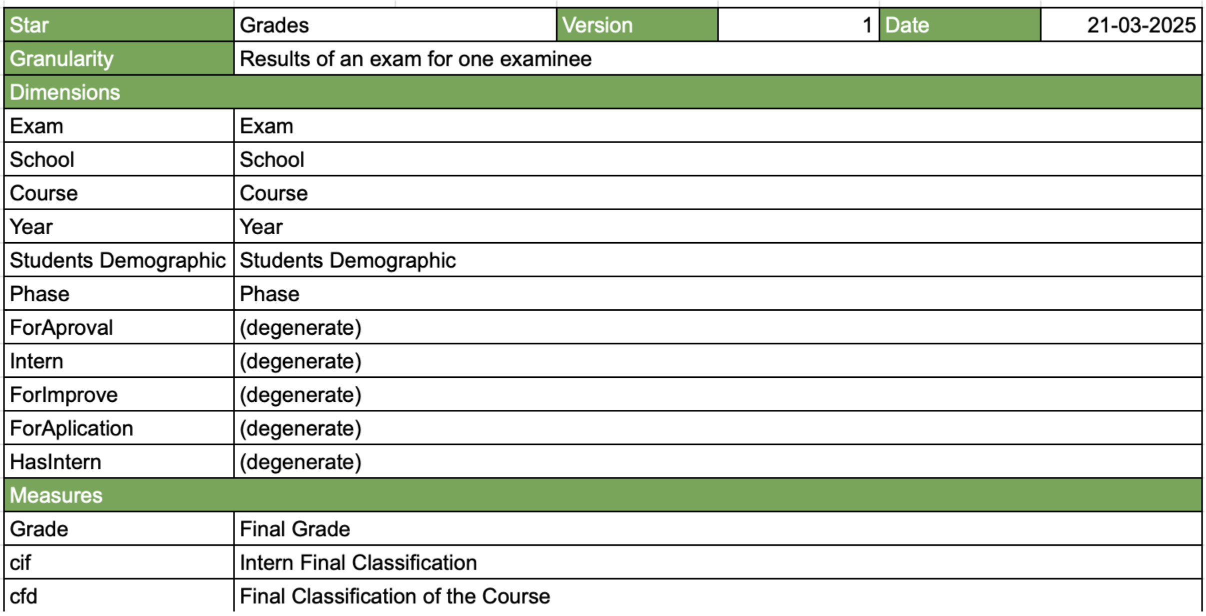
Our data warehouse design features six fact tables, each serving a specific analytical purpose and providing different perspectives on the Portuguese national examination data. These fact tables form the core of our dimensional model, enabling comprehensive analysis of exam results across various dimensions.

2.2.1 Grades

This is our primary and most granular fact table, capturing individual examination results at the student level. With a granularity defined as "Results of an exam for one examinee," it contains approximately one million records from the 2013 and 2014 examinations. This table connects to six dimensions, which are Exam, School, Course, Year, Students Demographic, and Phase and includes several degenerate dimensions that capture specific attributes related to the exam purpose: *ForAproval*, *Intern*, *ForImprove*, *ForAplication*, *HasIntern*. The *Grades* includes three critical measures that are central to the evaluation process:

- Grade: The actual score received by the student

- cif (Intern Final Classification): The internal evaluation score
- cfd (Final Classification of the Course): The final course classification



Star	Grades	Version	1	Date	21-03-2025
Granularity	Results of an exam for one examinee				
Dimensions					
Exam	Exam				
School	School				
Course	Course				
Year	Year				
Students Demographic	Students Demographic				
Phase	Phase				
ForAproval	(degenerate)				
Intern	(degenerate)				
ForImprove	(degenerate)				
ForAplication	(degenerate)				
HasIntern	(degenerate)				
Measures					
Grade	Final Grade				
cif	Intern Final Classification				
cfd	Final Classification of the Course				

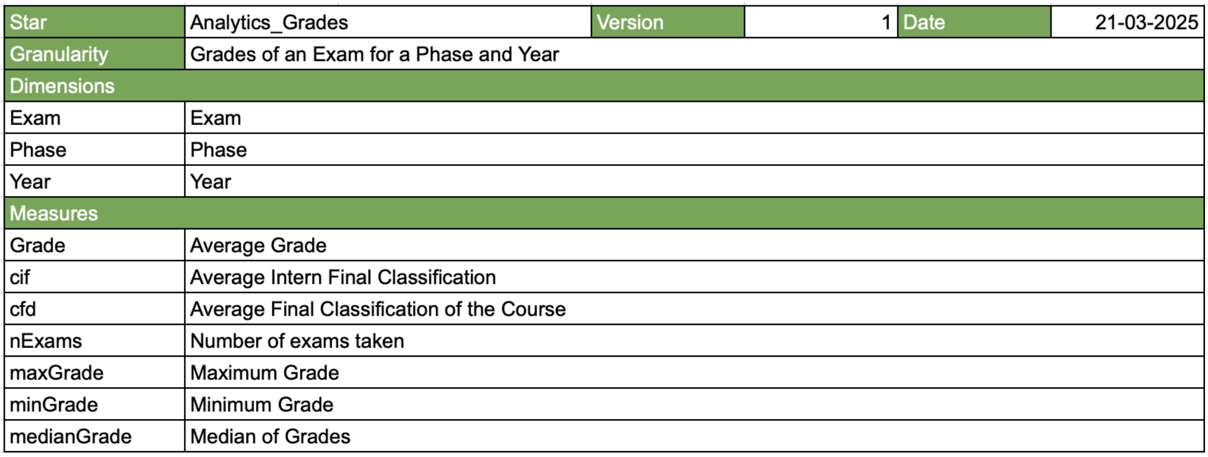
2.2.2 Analytics_Grades

This fact table aggregates examination results by Exam, Phase, and Year, providing a temporal view of exam performance. With a granularity of "Grades of an Exam for a Phase and Year," it offers a higher-level perspective than the Grades table.

The table includes several important average measures and comprehensive statistical measures:

- Grade: Average Grade
- cif: Average Intern Final Classification
- cfd: Average Final Classification of the Course
- nExams: Number of exams taken
- maxGrade: Maximum Grade
- minGrade: Minimum Grade
- medianGrade: Median of Grades

This table is particularly useful to track performance trends across different examination phases and academic years.

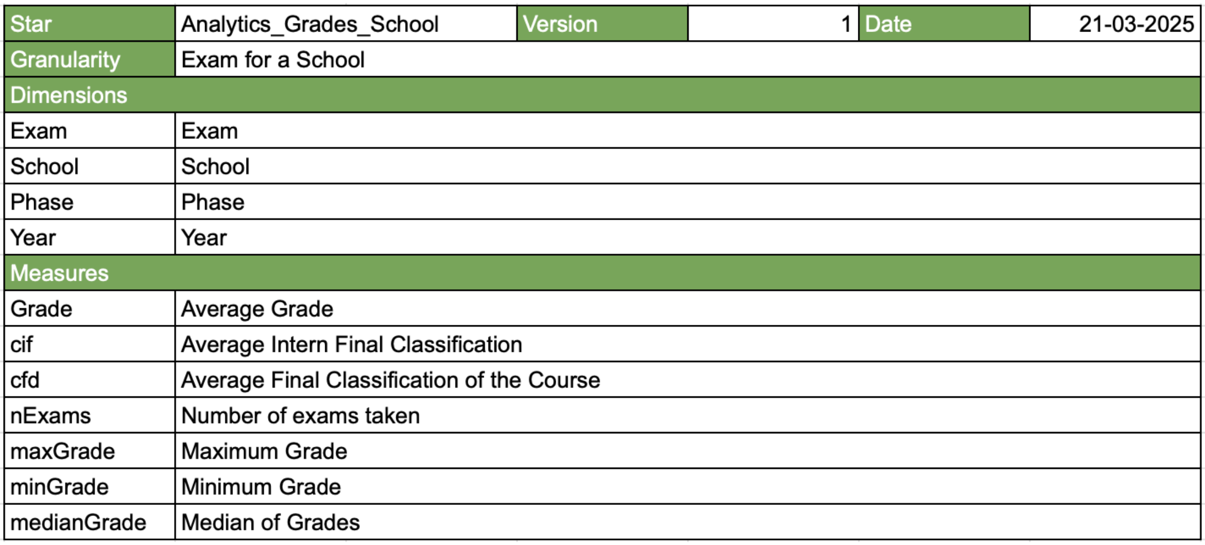


Star	Analytics_Grades	Version	1	Date	21-03-2025
Granularity	Grades of an Exam for a Phase and Year				
Dimensions					
Exam	Exam				
Phase	Phase				
Year	Year				
Measures					
Grade	Average Grade				
cif	Average Intern Final Classification				
cfd	Average Final Classification of the Course				
nExams	Number of exams taken				
maxGrade	Maximum Grade				
minGrade	Minimum Grade				
medianGrade	Median of Grades				

2.2.3 Analytics_Grades_School

This fact table focuses on district-level exam performance, with dimensions including Exam, District, Year, and Demographic Data. It enables comparative analysis across different districts, helping to identify high-performing and underperforming areas. The granularity of this table is “analytic Grades of Exam by School, Phase and Year.”

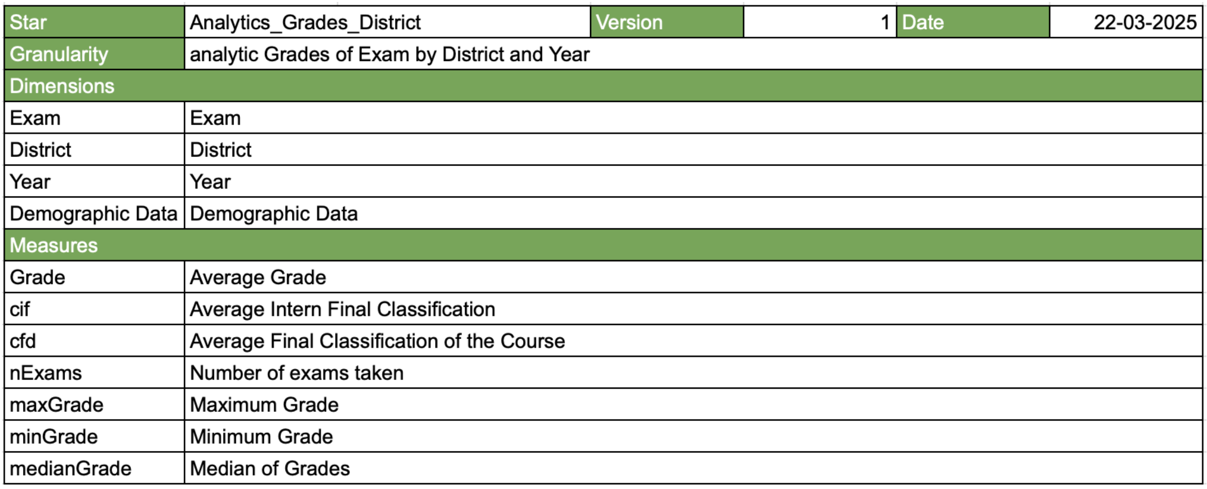
The statistical measures in this table (Grade, cif, cfd, nExams, maxGrade, minGrade, medianGrade) provide a comprehensive view of each district's performance. This table is particularly valuable to evaluate district effectiveness across different years.



Star	Analytics_Grades_School	Version	1	Date	21-03-2025
Granularity	Exam for a School				
Dimensions					
Exam	Exam				
School	School				
Phase	Phase				
Year	Year				
Measures					
Grade	Average Grade				
cif	Average Intern Final Classification				
cfd	Average Final Classification of the Course				
nExams	Number of exams taken				
maxGrade	Maximum Grade				
minGrade	Minimum Grade				
medianGrade	Median of Grades				

2.2.4 Analytics_Grades_District

With a granularity of "analytic Grades of Exam by District and Year," this fact table enables geographic analysis of examination results. By connecting to the Exam, District, Year, and Demographic Data dimensions, it allows for the exploration of regional performance patterns and disparities.

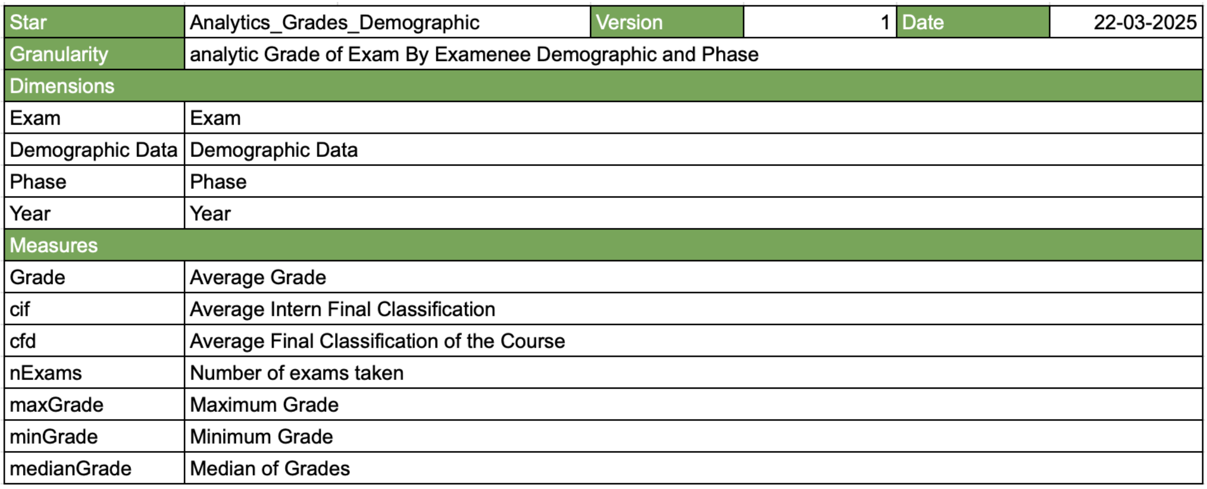


Star	Analytics_Grades_District	Version	1	Date	22-03-2025
Granularity	analytic Grades of Exam by District and Year				
Dimensions					
Exam	Exam				
District	District				
Year	Year				
Demographic Data	Demographic Data				
Measures					
Grade	Average Grade				
cif	Average Intern Final Classification				
cfd	Average Final Classification of the Course				
nExams	Number of exams taken				
maxGrade	Maximum Grade				
minGrade	Minimum Grade				
medianGrade	Median of Grades				

2.2.5 Analytics_Grades_Demographic

This fact table focuses on demographic analysis with a granularity of "analytic Grade of Exam By Examinee Demographic and Phase." By connecting to the Exam, Demographic Data, Phase, and Year dimensions, it enables the investigation of performance patterns across different demographic groups.

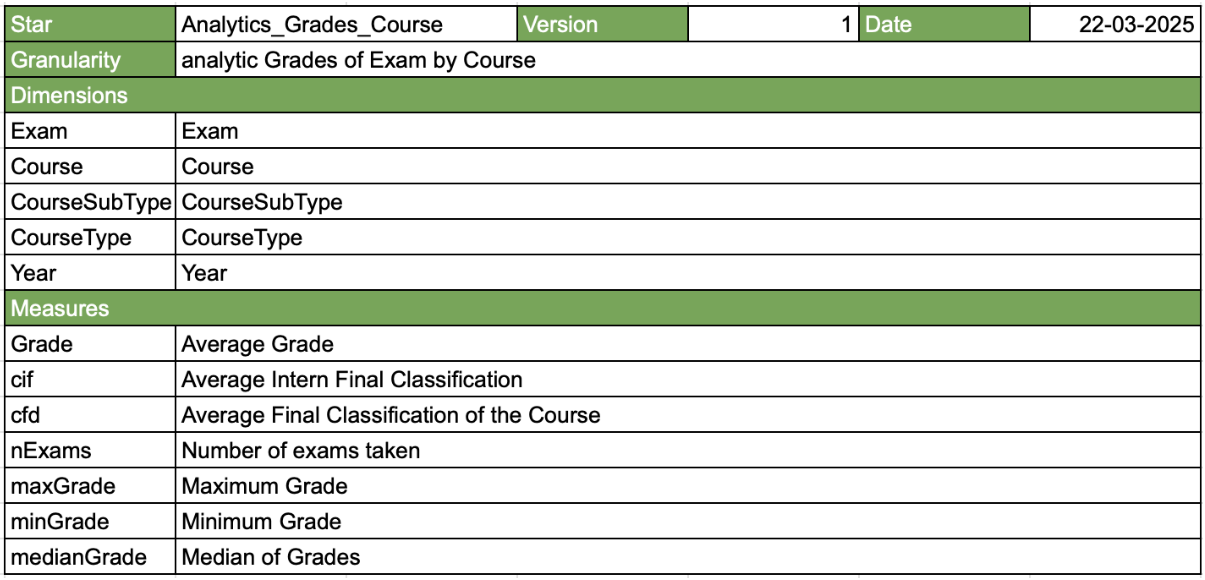
This table is particularly valuable to identify and address performance gaps between different demographic segments. The comprehensive statistical measures provide depth to this analysis, revealing not just average performance but also the range and distribution of scores.



Star	Analytics_Grades_Demographic	Version	1	Date	22-03-2025
Granularity	analytic Grade of Exam By Examenee Demographic and Phase				
Dimensions					
Exam	Exam				
Demographic Data	Demographic Data				
Phase	Phase				
Year	Year				
Measures					
Grade	Average Grade				
cif	Average Intern Final Classification				
cfd	Average Final Classification of the Course				
nExams	Number of exams taken				
maxGrade	Maximum Grade				
minGrade	Minimum Grade				
medianGrade	Median of Grades				

2.2.6 Analytics_Grades_Course

The final fact table in our design focuses on curriculum-based analysis, with a granularity of "analytic Grades of Exam by Course." By connecting to the Exam, Course, CourseSubType, CourseType, and Year dimensions, it enables detailed analysis of performance across different subject areas and course types.



Star	Analytics_Grades_Course	Version	1	Date	22-03-2025
Granularity	analytic Grades of Exam by Course				
Dimensions					
Exam	Exam				
Course	Course				
CourseSubType	CourseSubType				
CourseType	CourseType				
Year	Year				
Measures					
Grade	Average Grade				
cif	Average Intern Final Classification				
cfd	Average Final Classification of the Course				
nExams	Number of exams taken				
maxGrade	Maximum Grade				
minGrade	Minimum Grade				
medianGrade	Median of Grades				

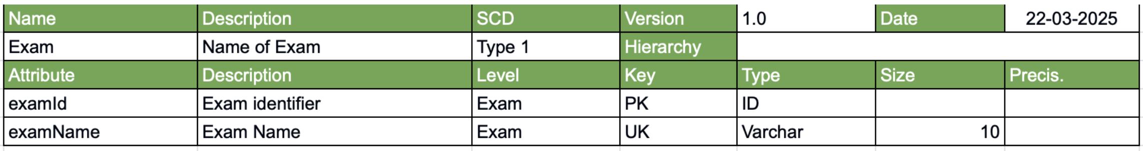
Together, these fact tables provide a comprehensive analytical framework for the Portuguese national examination system, enabling stakeholders at all levels to extract valuable insights from the examination data. Each table serves a specific analytical purpose while maintaining dimensional consistency through shared conformed dimensions, ensuring integrated analysis across the entire data warehouse.

2.3 Dimensions Dictionary

Our data warehouse design features nine well-defined dimensions that enable multidimensional analysis of Portuguese national examination results.

2.3.1 Exam Dimension

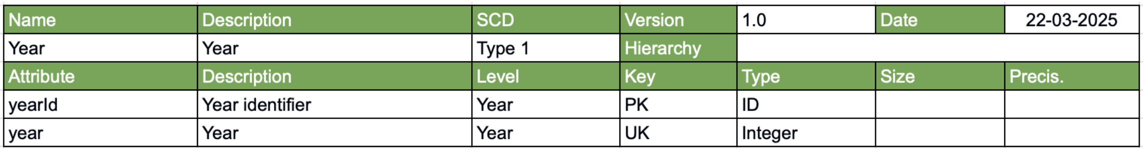
This dimension represents the specific examination subjects offered in the Portuguese national examination data. It is used in all fact tables and serves as a primary analytical axis, allowing performance analysis by subject area.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Exam	Name of Exam	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
examId	Exam identifier	Exam	PK	ID		
examName	Exam Name	Exam	UK	Varchar	10	

2.3.2 Year Dimension

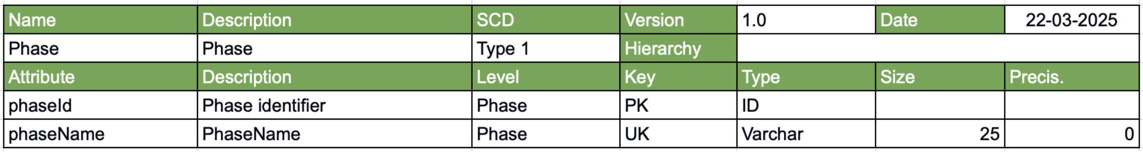
This temporal dimension represents the academic year in which examinations were conducted. It appears in all fact tables and enables trend analysis across different academic periods.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Year	Year	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
yearId	Year identifier	Year	PK	ID		
year	Year	Year	UK	Integer		

2.3.3 Phase Dimension

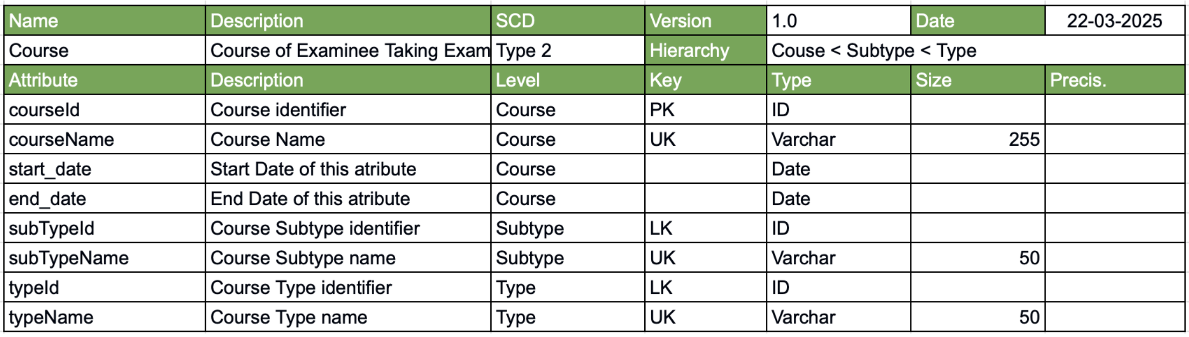
This dimension represents the examination phase (e.g., first phase, second phase) within the Portuguese examination. It allows analysis of performance differences between examination phases.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Phase	Phase	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
phaseId	Phase identifier	Phase	PK	ID		
phaseName	PhaseName	Phase	UK	Varchar	25	0

2.3.4 Course Dimension

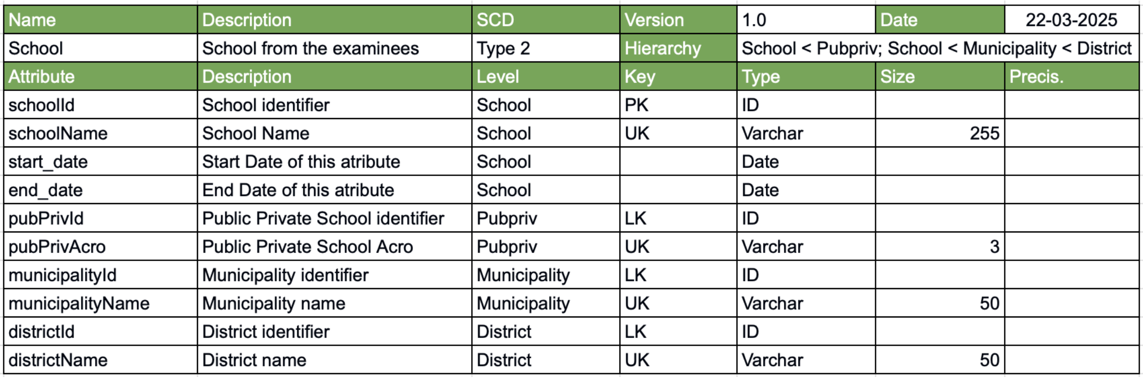
This dimension represents the educational courses that students are enrolled in when taking examinations. It enables curriculum-based analysis, showing how different courses perform across examinations.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Course	Course of Examinee Taking Exam	Type 2	Hierarchy	Couse < Subtype < Type		
Attribute	Description	Level	Key	Type	Size	Precis.
courseId	Course identifier	Course	PK	ID		
courseName	Course Name	Course	UK	Varchar	255	
start_date	Start Date of this atribute	Course		Date		
end_date	End Date of this atribute	Course		Date		
subTypeId	Course Subtype identifier	Subtype	LK	ID		
subTypeName	Course Subtype name	Subtype	UK	Varchar	50	
typeId	Course Type identifier	Type	LK	ID		
typeName	Course Type name	Type	UK	Varchar	50	

2.3.5 School Dimension

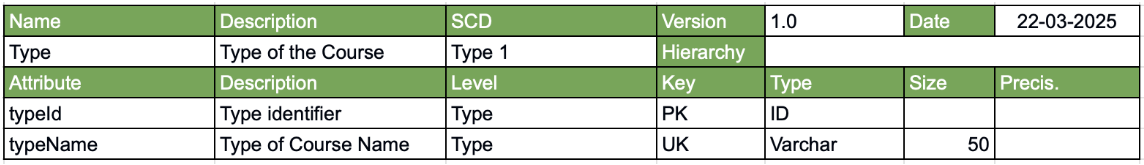
This dimension represents the educational institutions where students take examinations. It enables institutional analysis, allowing comparison of school performance.



Name	Description	SCD	Version	1.0	Date	22-03-2025
School	School from the examinees	Type 2	Hierarchy	School < Pubpriv; School < Municipality < District		
Attribute	Description	Level	Key	Type	Size	Precis.
schoolId	School identifier	School	PK	ID		
schoolName	School Name	School	UK	Varchar	255	
start_date	Start Date of this atribute	School		Date		
end_date	End Date of this atribute	School		Date		
pubPrivId	Public Private School identifier	Pubpriv	LK	ID		
pubPrivAcro	Public Private School Acro	Pubpriv	UK	Varchar	3	
municipalityId	Municipality identifier	Municipality	LK	ID		
municipalityName	Municipality name	Municipality	UK	Varchar	50	
districtId	District identifier	District	LK	ID		
districtName	District name	District	UK	Varchar	50	

2.3.6 Type Dimension

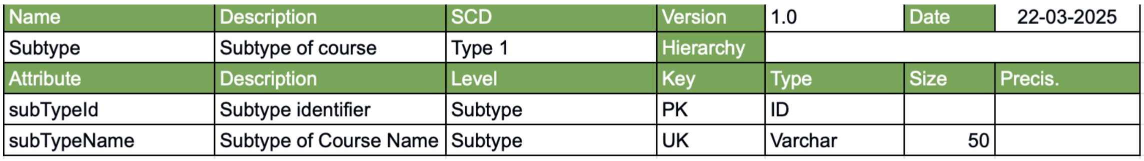
This dimension represents the top level categorization of courses in the educational system. It provides the highest level of course categorization.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Type	Type of the Course	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
typeId	Type identifier	Type	PK	ID		
typeName	Type of Course Name	Type	UK	Varchar	50	

2.3.7 Subtype Dimension

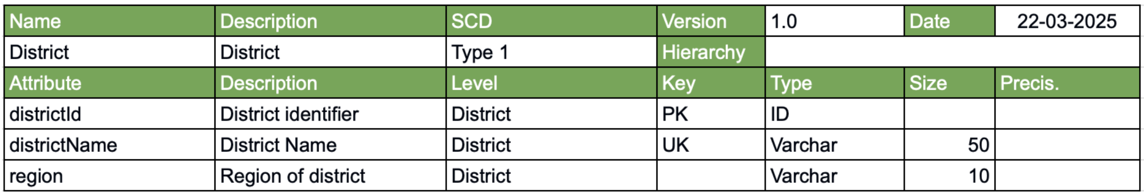
This dimension represents the mid-level categorization of courses, between course and type. This dimension enables more granular curriculum analysis than Type alone, allowing for medium-level educational program comparison.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Subtype	Subtype of course	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
subTypeId	Subtype identifier	Subtype	PK	ID		
subTypeName	Subtype of Course Name	Subtype	UK	Varchar	50	

2.3.8 District Dimension

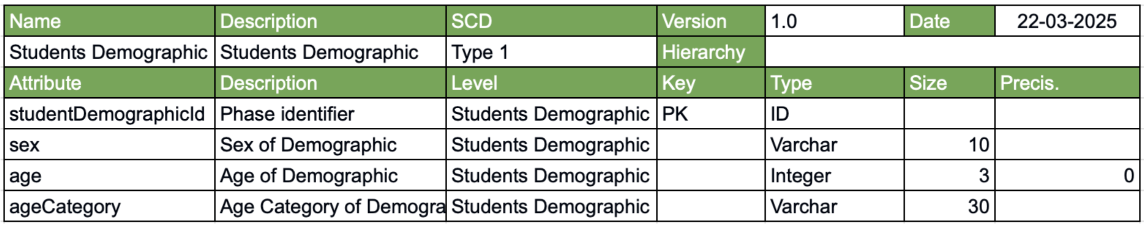
This dimension represents the districts of Portugal where schools are located. It supports regional analysis of examination performance, helping identify geographic patterns and disparities.



Name	Description	SCD	Version	1.0	Date	22-03-2025
District	District	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
districtId	District identifier	District	PK	ID		
districtName	District Name	District	UK	Varchar	50	
region	Region of district	District		Varchar	10	

2.3.9 Students Demographic Dimension

This dimension represents demographic attributes of the students taking examinations. It enables demographic analysis of examination performance, helping identify patterns related to gender and age groups.



Name	Description	SCD	Version	1.0	Date	22-03-2025
Students Demographic	Students Demographic	Type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precis.
studentDemographicId	Phase identifier	Students Demographic	PK	ID		
sex	Sex of Demographic	Students Demographic		Varchar	10	
age	Age of Demographic	Students Demographic		Integer	3	0
ageCategory	Age Category of Demogra	Students Demographic		Varchar	30	

3. Dimensional Data Model

The dimensional data model represents the core of our data warehouse design. Our model follows a star schema design pattern, with multiple fact tables sharing conformed dimensions in what is technically known as a "constellation schema." This architecture has been chosen to optimize query performance, simplify end-user access, and support a wide range of analytical perspectives.

3.1 Overall Architecture

As shown in the diagram (Figure 2 in Appendix), our dimensional model consists of six fact tables (highlighted in yellow) and nine dimension tables (represented in white). The model is organized around the central concept of examination results, with each fact table representing a different analytical perspective. Detailed descriptions of the fact tables and dimension tables, including their definitions, have been provided in Chapters 2.2 and 2.3, respectively, where the facts dictionary and dimensions dictionary are fully explained.

3.2 Star Schema Design

Each fact table in our model serves as the central point of its own star schema, with the dimension tables extending from it. For example, the Grades fact table connects directly to six dimensions (*dim_exam*, *dim_school*, *dim_course*, *dim_year*, *dim_studentDemographic*, and *dim_phase*). This star pattern optimizes query performance by minimizing the number of joins required for analysis while providing intuitive navigation paths for users.

The Dimensional Model supports several hierarchical relationships:

1. Geographic Hierarchy: *dim_school* → *dim_district* (with municipality as an intermediate level)
2. Curriculum Hierarchy: *dim_course* → *dim_subType* → *dim_type*
3. Demographic Grouping: *age* → *ageCategory* in *dim_studentDemographic*

These hierarchies enable drill-down and roll-up operations, allowing users to navigate between different levels of detail based on their analytical needs.

3.3 Fact Tables and Measures

Our fact tables contains several statistics measures:

- grade/averageGrade Examination scores
- cif/averageCif: Internal final classification
- cfd/averageCfd: Final course classification
- nExams: Count of examinations
- maxGrade/minGrade: Highest and lowest grades
- medianGrade: Median grade value

The Grades fact table contains individual exam records with additive measures (grade, cif, cfd), while the Analytics Fact tables contain pre-aggregated statistics with semi-additive measures (averages, medians).

3.4 Dimension Design Considerations

Our dimension tables implement several important design patterns:

- Slowly Changing Dimensions (Type 2): The *dim_school* and *dim_course* dimensions include *start_date* and *end_date* attributes to track changes over time, preserving historical accuracy.
- Descriptive Attributes: Each dimension includes both technical keys and descriptive attributes to facilitate intuitive analysis and reporting.
- Hierarchical Support: Several dimensions, such as *dim_course* and *dim_school*, include hierarchical relationships, enabling drill-down and roll-up operations.
- Conformed Dimensions: Dimensions like *dim_exam* and *dim_year* are shared across all fact tables, ensuring consistent analysis throughout the data warehouse.
- Granularity Consistency: The dimensions have been carefully designed to match the granularity of the related fact tables, ensuring accurate aggregation.

4. ETL and Data Sources

The primary data source for this project is the Portuguese National Examination database ("Exames Nacionais do Ensino Secundário" or ENES), obtained from the Portuguese Ministry of Education's Directorate-General for Education (DGE) statistical reports website (https://www.dge.mec.pt/relatoriosestatisticas-0). This repository contains comprehensive data on national examinations conducted in Portugal, providing valuable insights into student performance across different subjects, schools, and regions.

The dataset contains results from all Portuguese national examinations, but this project specifically focuses on data from the years 2013 and 2014, encompassing approximately one million examination records, including detailed information on:

- Examination subjects and phases
- Student demographics (age, gender)

- Educational institutions (schools, municipalities, districts)
- Course types and subtypes
- Examination purposes (for approval, improvement, university admission)
- Grades and classifications (exam grade, internal classification, final course classification)

4.1 Extraction

The extraction process involved several steps to acquire the raw data from its source and prepare it for transformation:

1. **Data Download:** The examination datasets were downloaded from the Portuguese Ministry of Education's statistical portal in their original Access format, then transformed into CSV.
2. **Initial Database Creation:** The raw data was imported into a PostgreSQL database hosted on a PHPMyAdmin server (dbm.fe.up.pt), within the 'enes_2013' schema. The import preserved the original operational structure, including tables such as:
    - *tblhomologa*: The main table containing examination records
    - *tblescolas*: Schools information
    - *tblexames*: Examination definitions
    - *tblcursos*: Course information
    - *tblcursostipos*: Course types
    - *tblcursossub*: Course subtypes
    - *tblcodsdistrito*: District reference data
    - *tblcodsconcelho*: Municipality reference data
3. **Schema Analysis:** The operational schema was analyzed to understand entity relationships, primary and foreign keys, and data quality issues (as shown in Figure 1 in the Appendix).

This extraction phase maintained the original structure and values of the source data, preserving it within the operational database before any transformations were applied.

4.2 Transformation

The transformation phase involved several key processes to convert the operational data into a dimensional structure optimized for analytical queries:

1. **Dimensional Modeling:** The operational schema was redesigned into a star schema (Figure 2 in Appendix) with dimension and fact tables, creating clear separation between descriptive attributes (dimensions) and measures (facts).
2. **Data Type Conversions:** We converted examination grades and classifications from their source formats to appropriate numerical types (integers for grades, floating-point numbers for classifications).
3. **Derived Attributes:** New attributes were created to enhance analytical capabilities, such as:
   - Age categories in the student demographic dimension (grouping ages into meaningful ranges like "16-18 years", "19-21 years", etc.)
   - Regional classification for districts (grouping districts into Norte, Centro, Sul, and Regiões Autónomas)
   - Public/Private school classification (converting text codes to numeric identifiers)
4. **Hierarchical Structure:** Hierarchical relationships were explicitly modeled:
   - Geographic hierarchy (School → Municipality → District → Region)
   - Course hierarchy (Course → Subtype → Type)
5. **Statistical Aggregations:** For each analytical perspective, statistical measures were calculated:
   - Average grades and classifications
   - Count of examinations
   - Maximum and minimum grades
   - Median grade values
6. **Degenerate Dimension Transformation:** Boolean flags were created from categorical attributes in the source data to simplify filtering (forAproval, intern, forImprove, hasAplication, hasIntern).

These transformations substantially enhanced the analytical value of the data while maintaining its semantic integrity, creating a dimensional model tailored to educational performance analysis.

4.3 Loading

The loading phase populated the dimensional model (named 'data_warehouse' in PostgreSQL) with both atomic and pre-aggregated data according to a carefully sequenced process:

1. **Preparation of the Loading Structure:** A BPMN 2.0 ETL graph was created (Figure 3 in Appendix) to systematically define the sequence of table loading, ensuring an efficient and structured data integration process.
2. **Dimension Loading:** Dimensions were loaded first to establish the reference framework, extracting unique values from the source data and applying transformations to create the dimensional attributes.
3. **Atomic Fact Table Loading:** The primary fact table (grades) was loaded with individual examination records, linking to the appropriate dimension keys and including all relevant measures.
4. **Aggregated Fact Tables Creation:** Five aggregated fact tables—*average_grades, average_grades_district, average_grades_school, average_grades_demographic,* and *average_grades_course*—were designed and populated with pre-calculated statistical measures, each tailored to support distinct analytical perspectives.
5. **Statistical Calculations:** Advanced statistical functions were employed to compute aggregated measures, including averages, counts, maximum and minimum values, and medians.
6. **Referential Integrity:** Foreign key relationships were established between fact and dimension tables to maintain data integrity and enable efficient joins during analysis.

This ETL process transformed approximately one million examination records from a relational model into a dimensionally modeled data warehouse, enabling effective multidimensional analysis of Portuguese national exams.

5. Query and Data Analysis

This chapter analyzes the efficiency and performance of queries in the two different data models: the relational and the dimensional model. The comparison is based on query execution time and the impact of data organization on performance.

5.1 Query Performance Comparison

This section presents a comparison of query performance between relational and dimensional models for different types of data analysis.

5.1.1 Average Exam Grade by Year and Subject

Relational Model: 474.596 ms

```
SELECT e.descr AS exam_name, h.ano AS year, AVG(h.nota) AS avg_grade
FROM enes_2013.tblhomologa h
JOIN enes_2013.tblexames e ON h.exame = e.exame
GROUP BY e.descr, h.ano
ORDER BY h.ano, avg_grade DESC;
```

Dimensional Model: 2.192 ms

```
SELECT e.examName, y.year, a.averageGrade
FROM data_warehouse.analytics_grades a
JOIN data_warehouse.dim_exam e ON a.examId = e.examId
JOIN data_warehouse.dim_year y ON a.yearId = y.yearId
ORDER BY y.year, a.averageGrade DESC;
```

- **Key Differences:**
  - The relational model requires scanning all records and grouping dynamically, leading to slower execution.
  - The dimensional model uses pre-aggregated data, making it significantly faster.

5.1.2 Exam Performance by School Type (Public vs. Private)

Relational Model: 378.055 ms

```
      SELECT p.descr AS school_type, AVG(h.nota) AS avg_grade
      FROM enes_2013.tblhomologa h
      JOIN enes_2013.tblescolas s ON h.escola = s.escola
      JOIN enes_2013.tblcodspubpriv p ON s.pubpriv = p.pubpriv
      GROUP BY p.descr;
```

Dimensional Model: 12.631 ms

```
      SELECT s.pubPrivAcro AS school_type, AVG(g.averageGrade) AS
avg_grade
      FROM data_warehouse.analytics_grades_school g
      JOIN data_warehouse.dim_school s ON g.schoolId = s.schoolId
      GROUP BY s.pubPrivAcro;
```

- **Key Differences:**
    - The relational model requires text-based filtering and multiple joins, slowing down performance.
    - The dimensional model uses numeric keys and pre-aggregated data, leading to faster execution.

5.1.3 Regional Performance Trends Over Time

Relational Model: 398.463 ms

```
      SELECT d.distrito AS district, h.ano AS year, AVG(h.nota) AS
avg_grade
      FROM enes_2013.tblhomologa h
      JOIN enes_2013.tblescolas s ON h.escola = s.escola
      JOIN enes_2013.tblcodsdistrito d ON s.distrito = d.distrito
      GROUP BY d.distrito, h.ano
      ORDER BY h.ano, avg_grade DESC;
```

Dimensional Model: 176.549 ms

```
      SELECT d.districtName, y.year, a.averageGrade
      FROM data_warehouse.analytics_grades_district a
```

```
JOIN data_warehouse.dim_district d ON a.districtId = d.districtId
JOIN data_warehouse.dim_year y ON a.yearId = y.yearId
ORDER BY y.year, a.averageGrade DESC;
```

- **Key Differences**:
  - The relational model requires multiple joins across district and school tables.
  - The dimensional model encodes the geographic hierarchy, reducing query complexity and improving speed.

5.1.4 Compare Average Grades by District

Relational Model: 499.223 ms

```
SELECT d.descr AS district, AVG(h.class_exam::INT) AS avg_grade
FROM enes_2013.tblhomologa h
JOIN enes_2013.tblescolas e ON h.escola = e.escola
JOIN enes_2013.tblcodsdistrito d ON e.distrito = d.distrito
GROUP BY d.descr
ORDER BY avg_grade DESC;
```

Dimensional Model: 8.724 ms

```
SELECT d.districtName, AVG(a.averageGrade) AS avg_grade
FROM data_warehouse.analytics_grades_district a
JOIN data_warehouse.dim_district d ON a.districtId = d.districtId
GROUP BY d.districtName
ORDER BY avg_grade DESC;
```

- **Key Differences**:
  - The pre-aggregated data of the dimensional model eliminates the need for runtime computation.

5.1.5 Compare School Performance

Relational Model: 380.019 ms

```
SELECT e.descr AS school, AVG(h.class_exam::INT) AS avg_grade
FROM enes_2013.tblhomologa h
JOIN enes_2013.tblescolas e ON h.escola = e.escola
GROUP BY e.descr
```

```
ORDER BY avg_grade DESC;
```

Dimensional Model: 6.182 ms

```
SELECT s.schoolName, AVG(a.averageGrade) AS avg_grade
FROM data_warehouse.average_grades_school a
JOIN data_warehouse.dim_school s ON a.schoolId = s.schoolId
GROUP BY s.schoolName
ORDER BY avg_grade DESC;
```

- **Key Differences**
    - The dimensional model utilizes pre-aggregated average grades, reducing runtime calculations.
    - Integer-based keys improve join performance.

5.2 Results and Analysis

The comparative analysis confirms that queries executed in the dimensional model consistently outperform those in the relational model. The primary factors contributing to improved performance include:

- **Pre-aggregated Data:** The dimensional model stores pre-computed statistical measures, eliminating runtime calculations.
- **Optimized Indexing and Partitioning:** Efficient use of indexing speeds up query execution.
- **Reduced Joins:** The relational model requires multiple joins, while the dimensional model minimizes these operations by integrating conformed dimensions.

These findings highlight the **superior efficiency** of dimensional modeling for analytical workloads, where query performance is critical.

6. Visualizations

To leverage the analytical potential of our dimensional data warehouse, we developed a comprehensive set of interactive dashboards using Microsoft Power BI. These visualizations transform the raw examination data into actionable insights, enabling the exploration of performance patterns across multiple dimensions.

6.1 Dashboard Architecture

Our visualization solution consists of three interconnected dashboards, each focusing on a different analytical perspective:

1. **General Dashboard:** Provides a high-level overview of examination performance across subjects and phases
2. **Schools Dashboard:** Enables detailed analysis of institutional performance with ranking capabilities
3. **Districts Dashboard:** Offers geographic visualization of regional performance patterns

These dashboards are designed to work together as an integrated analytical solution while allowing users to focus on their specific areas of interest.

6.2 General Dashboard Analysis

The General Dashboard (Figure 4 in Appendix) serves as the primary entry point for data exploration, presenting key performance indicators and subject-specific insights:

- **Key Performance Metrics:** The dashboard prominently displays critical metrics including total examination count (899,000), average exam grade (90.82), average internal classification (13.43), and average final classification (10.80).
- **Phase Comparison Table:** The tabular view in the center provides detailed subject performance across examination phases, revealing interesting patterns. For example, Mathematics B shows a significant performance drop between phase 1 (76.72) and phase 2 (64.57), suggesting either increased difficulty in the second phase or different student demographics.
- **Subject Performance Chart:** The bar chart at the bottom ranks subjects by average performance, showing PLNM Intermedio with the highest average scores (approximately 148) and Mathematics B with significantly lower performance. This visualization immediately highlights subject areas that may require curricular attention.

- **Multi-dimensional Filtering**: The right panel provides filtering capabilities based on examination purpose (Aprovação, Melhoria, Ingresso), student status , and year (2013, 2014), enabling targeted analysis for specific student segments.

This dashboard effectively leverages our *average_grades* fact table, which provides pre-aggregated metrics by exam, phase, and year, enabling instantaneous response to complex analytical queries that would be computationally intensive with the original operational database.

6.3 Schools Dashboard Analysis

The Schools Dashboard (Figure 5 in Appendix) focuses on institutional performance comparison, providing valuable insights for educational administration:

- **School Performance Ranking:** The horizontal bar chart presents the top-performing schools based on average examination grades. Private institutions like Colégio Nossa Senhora do Rosário and Colégio Arautos do Evangelho demonstrate consistently higher performance, providing an opportunity for investigating effective educational practices.
- **Hierarchical Filtering:** The right panel allows filtering by year, region/district, specific exam, phase, and demographic characteristics. This multi-level filtering capability directly leverages our dimensional hierarchies (School → District → Region) and conformed dimensions across fact tables.
- **School Performance Analysis:** The dashboard maintains the same KPIs as the general view but contextualizes them within the selected filters, allowing for comparative analysis between specific schools and overall national performance.

This visualization primarily utilizes our *average_grades_school* fact table, which contains pre-aggregated metrics at the school level. The star schema design enables efficient drill-down from regional performance to specific school results without complex query logic.

6.4 Districts Dashboard Analysis

The Districts Dashboard (Figure 6 in Appendix) provides geographic contextualization of examination performance, revealing regional patterns and disparities:

- **Geographic Visualization:** The map visualization on the left displays districts with circle size indicating examination volume. This immediately highlights educational centers like Porto, Lisbon, and Braga where large numbers of examinations are conducted.
- **District Performance Trend:** The line chart shows average grades by district in descending order, revealing a clear performance gradient across regions. Districts like Aveiro, Coimbra, and Porto consistently outperform others, while Estrangeiro (foreign locations) shows the lowest performance.
- **Detailed Metrics Table:** The table at the bottom provides detailed performance metrics for each district, enabling precise comparison of average grades and other key indicators.
- **Regional Pattern Analysis:** The combined visualizations reveal clear north-south and urban-rural performance patterns.

This dashboard primarily leverages our *average_grades_district* fact table, demonstrating the value of our geographic dimension hierarchy and regional attribute derivation during the ETL process.

6.5 Benefits of Dimensional Model for Visualization and Data Analysis

These powerful visualization capabilities directly demonstrate the advantages of our dimensional data warehouse design:

1. **Query Performance:** The fact tables enable instantaneous response to complex analytical queries, even with hundreds of thousands of underlying examination records.
2. **Consistent Analysis:** Conformed dimensions ensure consistent analysis across different analytical perspectives (general, school, district), maintaining analytical integrity.
3. **Derived Attributes:** Calculated attributes like age categories, regional classifications, and public/private designations provide immediate analytical value without additional transformation during visualization.

4. **Hierarchical Navigation:** The explicitly modeled hierarchies support intuitive drill-down capabilities, enhancing the exploratory analytical experience.

The visualizations presented in these dashboards would be significantly more difficult to develop and much less responsive if built directly on the original operational database. The dimensional model's optimization for analytical queries enables rich, interactive visualizations and analysis that provide immediate insights into Portuguese national examination performance.

7. Conclusions

7.1 Pros and Cons of Data Warehouses vs Operational Databases

Our project of transforming the Portuguese national examination data from an operational database into a dimensional data warehouse has provided valuable insights into the strengths and limitations of both approaches.

7.1.1 Advantages of Data Warehouses

1. **Query Performance:** The star schema design dramatically improves performance for analytical queries. Complex analyses that would require numerous joins and calculations in the operational database execute much faster in our dimensional model.
2. **Analytical Flexibility:** Our data warehouse design with multiple fact tables supports various analytical perspectives (by school, district, course, demographics) while maintaining dimensional consistency.
3. **Business User Accessibility:** The dimensional model is more intuitive for non-technical users to understand and navigate.
4. **Aggregation Benefits:** Pre-aggregated fact tables (like our *Analytics_Grades* tables) provide immediate access to common metrics without needing to recalculate them for every query.
5. **Historical Analysis:** The slowly changing dimension design (type 2) for schools and courses preserves historical context, facilitating accurate trend analysis across academic years.

6. **Enhanced Metadata:** Our dimensional model captures meaningful business definitions and hierarchies (geographic, curricular) that were merely implicit in the operational model.

7.1.2 Limitations of Data Warehouses

1. **Increased Complexity:** Implementing the ETL process required significant development effort, including complex transformations and data validation.
2. **Storage Redundancy:** The dimensional model introduces some redundancy, particularly in pre-aggregated fact tables, increasing storage requirements.
3. **Update Latency:** Unlike the operational database, our data warehouse follows a batch update model, creating some delay between data creation and analytical availability.

7.1.3 When Operational Databases Excel

1. **Transactional Processing:** The operational database remains superior for recording individual examination results efficiently.
2. **Data Entry and Updates:** For corrections or updates to individual examination records, the operational database provides simpler, atomic transaction capabilities.
3. **Real-time Requirements:** When immediate access to newly entered data is critical, the operational database avoids the ETL delay.
4. **Storage Efficiency:** The normalized structure minimizes redundancy, conserving storage resources.

In the context of Portuguese national examination analysis, the benefits of the dimensional data warehouse approach clearly outweigh its limitations. The significant improvements in analytical capabilities, query performance, and accessibility for education stakeholders justify the additional implementation complexity and resource requirements.

7.2 Future Developments

The field of data warehousing continues to evolve rapidly, with several emerging technologies and approaches that could enhance future iterations of educational data warehousing:

7.2.1 Cloud-Native Data Warehousing

Cloud platforms like Snowflake, Google BigQuery, and Amazon Redshift are transforming the data warehousing landscape. Snowflake's architecture separates storage and compute resources, allowing educational institutions to scale based on actual usage patterns - particularly valuable for handling seasonal examination data processing peaks without over provisioning resources. These platforms also offer built-in features for data sharing between educational institutions and regulatory bodies while maintaining appropriate security boundaries.

7.2.2 Data Lakehouse Architecture

The emerging data lakehouse paradigm combines data lake flexibility with data warehouse performance. For educational data that spans structured examination results, semi-structured feedback forms, and unstructured assessment materials, a lakehouse approach could provide a more comprehensive analytical foundation. Technologies like Delta Lake and Apache Iceberg maintain data quality while accommodating diverse educational data types.

7.2.3 AI/ML Integration

Modern data warehousing platforms increasingly integrate with AI and machine learning capabilities. For educational data, this could enable predictive analytics (identifying students at risk based on examination patterns), anomaly detection (identifying unusual performance patterns that might indicate issues), and recommendation systems (suggesting personalized interventions based on performance data).

The Portuguese national examination data warehouse we've developed provides a solid foundation that can evolve alongside these technological advancements. As these technologies mature, they offer exciting possibilities for enhancing educational data analysis, ultimately supporting more data-driven decision-making in educational contexts and improving student outcomes.

8. References

Kimball, R. (2007). *The data warehouse lifecycle toolkit*. Wiley. ISBN: 9781118075043

David, G. S. T. *Lecture slides*. Moodle platform. Retrieved from https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=541169

9. Appendix

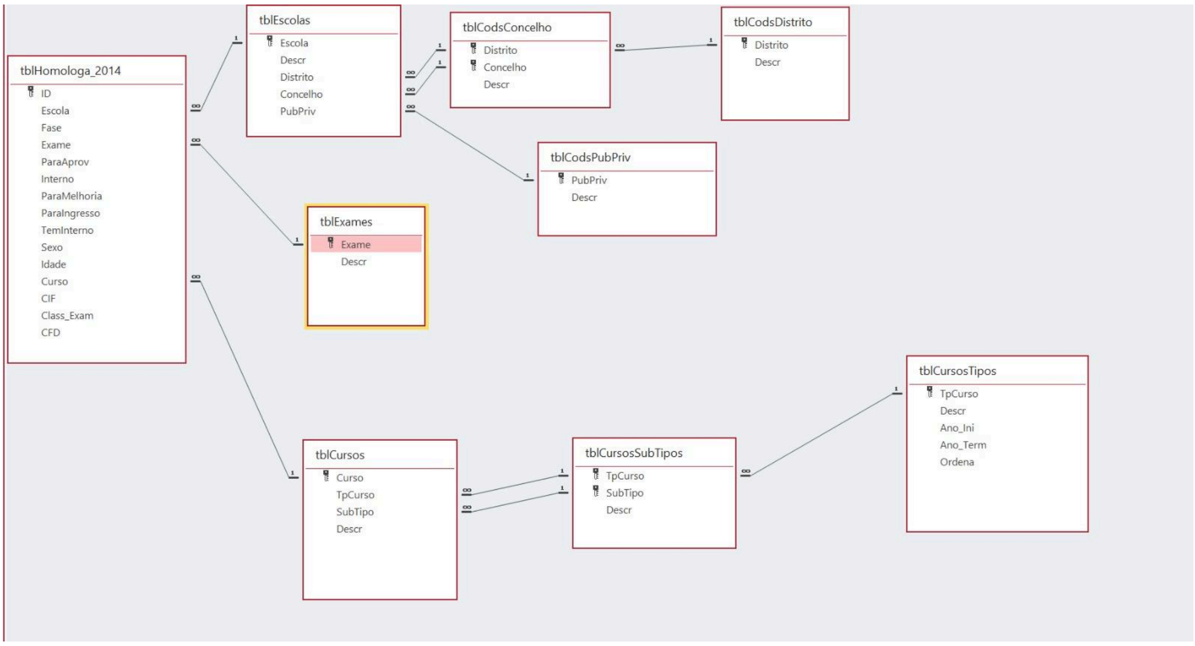


Figure 1- UML Operational Model

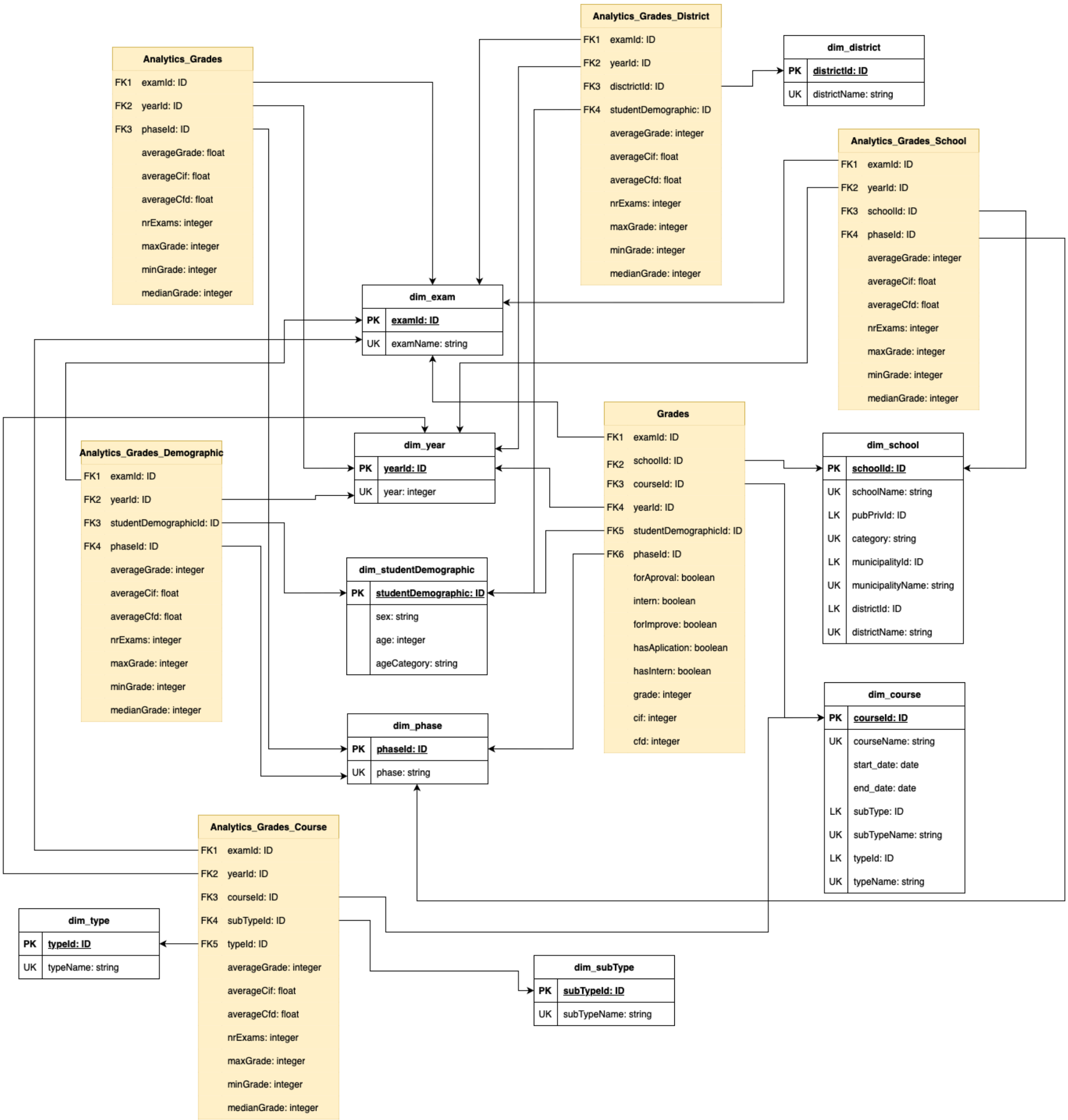


Figure 2 - UML Dimensional Model

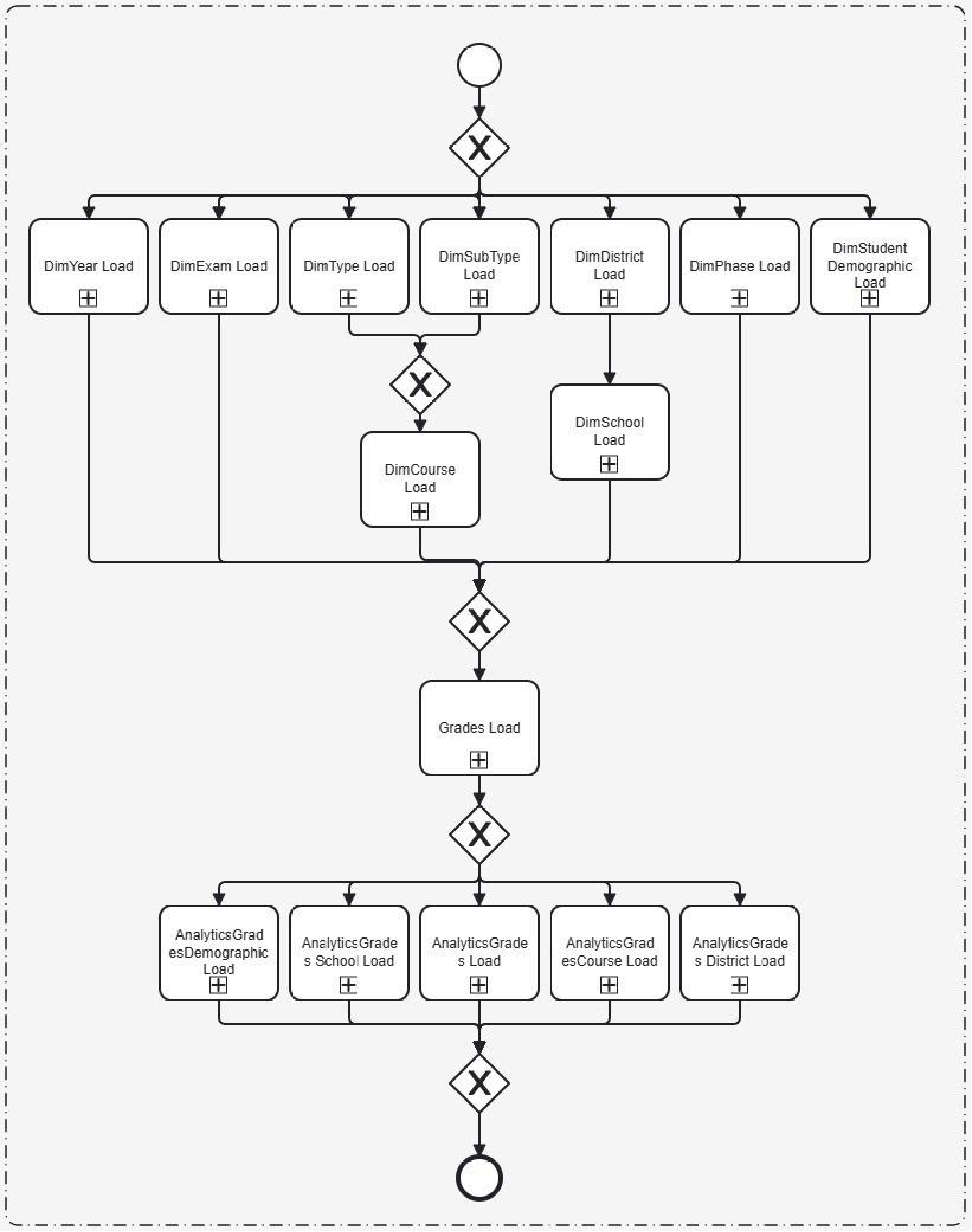
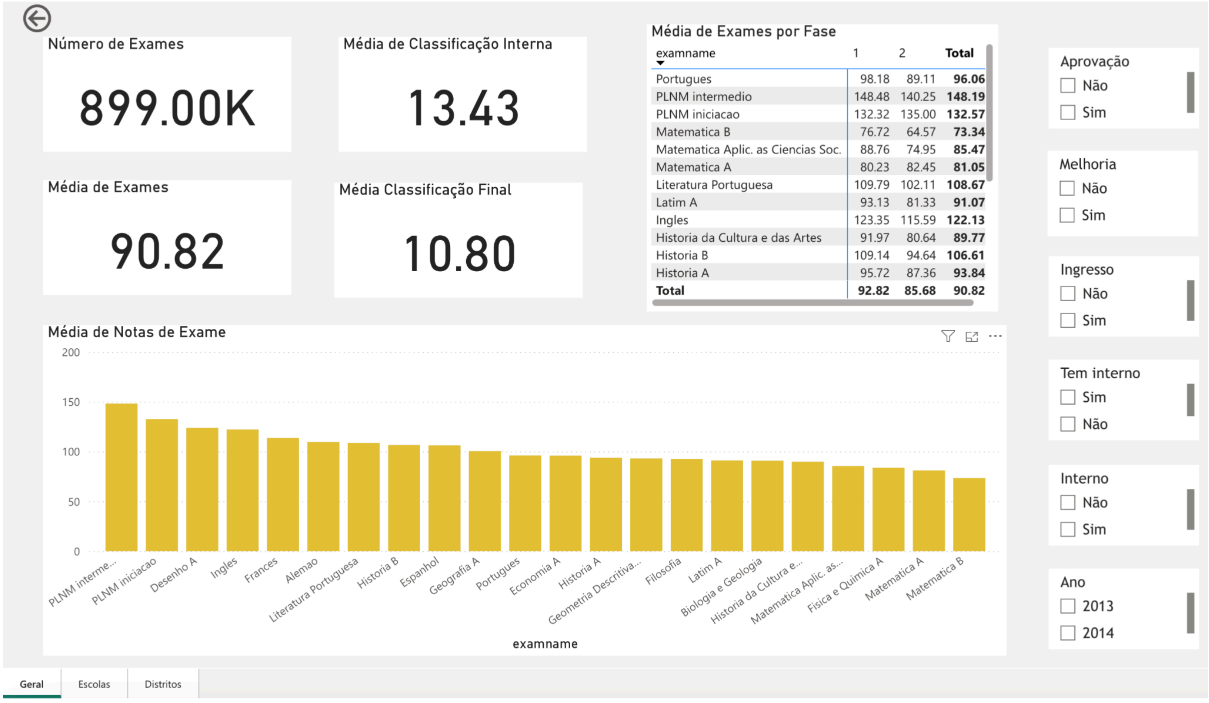

DimYear Load
DimExam Load
DimType Load
DimSubType Load
DimDistrict Load
DimPhase Load
DimStudent Demographic Load
DimCourse Load
DimSchool Load
Grades Load
AnalyticsGradesDemographic Load
AnalyticsGrades School Load
AnalyticsGrades Load
AnalyticsGradesCourse Load
AnalyticsGrades District Load

Figure 3 - BPMN 2.0 ETL



Número de Exames

899.00K

Média de Classificação Interna

13.43

Média de Exames

90.82

Média Classificação Final

10.80

Média de Exames por Fase

examname	1	2	Total
Portugues	98.18	89.11	**96.06**
PLNM intermedio	148.48	140.25	**148.19**
PLNM iniciacao	132.32	135.00	**132.57**
Matematica B	76.72	64.57	**73.34**
Matematica Aplic. as Ciencias Soc.	88.76	74.95	**85.47**
Matematica A	80.23	82.45	**81.05**
Literatura Portuguesa	109.79	102.11	**108.67**
Latim A	93.13	81.33	**91.07**
Ingles	123.35	115.59	**122.13**
Historia da Cultura e das Artes	91.97	80.64	**89.77**
Historia B	109.14	94.64	**106.61**
Historia A	95.72	87.36	**93.84**
Total	**92.82**	**85.68**	**90.82**

Aprovação
☐ Não
☐ Sim

Melhoria
☐ Não
☐ Sim

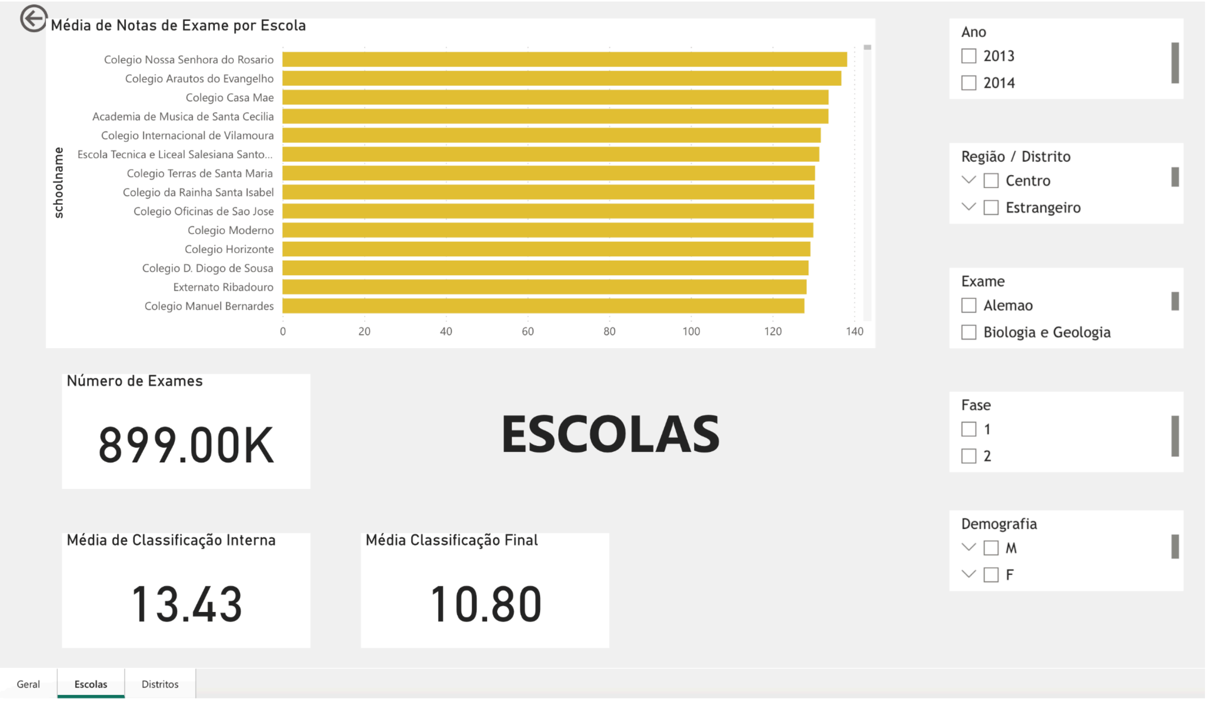
Ingresso
☐ Não
☐ Sim

Tem interno
☐ Sim
☐ Não

Interno
☐ Não
☐ Sim

Ano
☐ 2013
☐ 2014

Figure 4 - General Dashboard



Média de Notas de Exame por Escola
schoolname
Colegio Nossa Senhora do Rosario
Colegio Arautos do Evangelho
Colegio Casa Mae
Academia de Musica de Santa Cecilia
Colegio Internacional de Vilamoura
Escola Tecnica e Liceal Salesiana Santo...
Colegio Terras de Santa Maria
Colegio da Rainha Santa Isabel
Colegio Oficinas de Sao Jose
Colegio Moderno
Colegio Horizonte
Colegio D. Diogo de Sousa
Externato Ribadouro
Colegio Manuel Bernardes
0
20
40
60
80
100
120
140

Ano
☐ 2013
☐ 2014

Região / Distrito
☐ Centro
☐ Estrangeiro

Exame
☐ Alemao
☐ Biologia e Geologia

Fase
☐ 1
☐ 2

Demografia
☐ M
☐ F

Número de Exames

899.00K

ESCOLAS

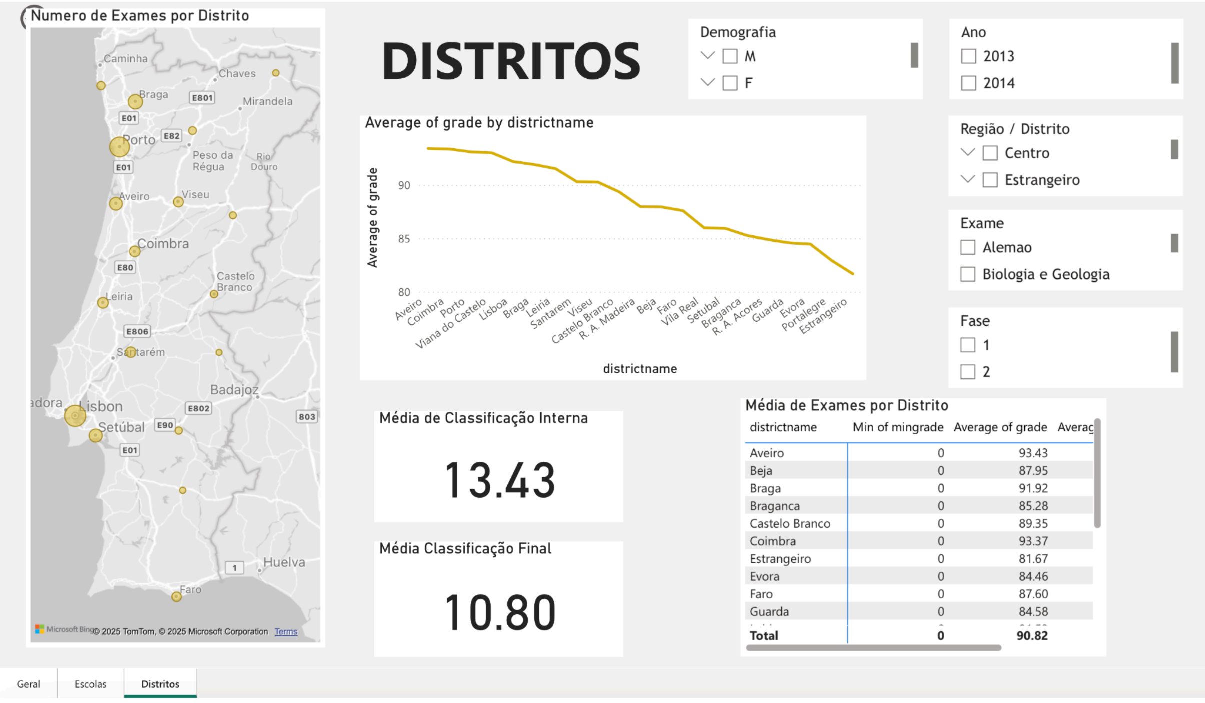
Média de Classificação Interna

13.43

Média Classificação Final

10.80

Figure 5 - School Dashboard



Numero de Exames por Distrito

DISTRITOS

Demografia

- □ M
- □ F

Ano

- □ 2013
- □ 2014

Região / Distrito

- □ Centro
- □ Estrangeiro

Exame

- □ Alemao
- □ Biologia e Geologia

Fase

- □ 1
- □ 2

Average of grade by districtname

Média de Classificação Interna

13.43

Média Classificação Final

10.80

Média de Exames por Distrito

districtname	Min of mingrade	Average of grade	Averag
Aveiro	0	93.43	
Beja	0	87.95	
Braga	0	91.92	
Braganca	0	85.28	
Castelo Branco	0	89.35	
Coimbra	0	93.37	
Estrangeiro	0	81.67	
Evora	0	84.46	
Faro	0	87.60	
Guarda	0	84.58	
Total	**0**	**90.82**	

Geral | Escolas | Distritos

Figure 6 - District Dashboard